



Universiteit
Leiden
Governance and Global Affairs

udp UNIVERSIDAD
DIEGO PORTALES



Validación de un estándar de referencia para medir toxicidad e incivilidad política en la esfera digital

Bastián González-Bustamante

✉ b.a.gonzalezbustamante@fgga.leidenuniv.nl

Sebastián Rivera

✉ sebastian.rivera@uci.edu

Presentación Proyecto FAE-UDP 2024
14 de julio de 2024

Contenidos

1. Proyecto FAE-UDP 2024
2. Datos y clasificación de toxicidad
3. Etiquetado humano
4. Trabajo futuro
5. Comentarios finales



Proyecto FAE-UDP 2024

Proyecto FAE-UDP 2024

Este proyecto tiene por objetivo crear y validar un estándar de referencia (**gold standard**) para un conjunto de algoritmos de aprendizaje automático y modelos de lenguaje que miden toxicidad e incivilidad política en la esfera digital.

Timetable

- Recolección y limpieza de datos
- Aplicación de primeros algoritmos clasificatorios
- Proceso de etiquetado humano
- Aplicación de etiquetado con LLMs (**OpenAI project**)
- Benchmarking

Datos y clasificación de toxicidad

Datos y clasificación de toxicidad

Eventos de protesta

-  Argentina, agosto 2020
 -  Brasil, mayo 2019
 -  Chile, octubre 2019
- ~5M de mensajes

Convención Constitucional

-  Interacciones con convencionales
- Dos años de data
- ~31M de mensajes

Clasificación de toxicidad

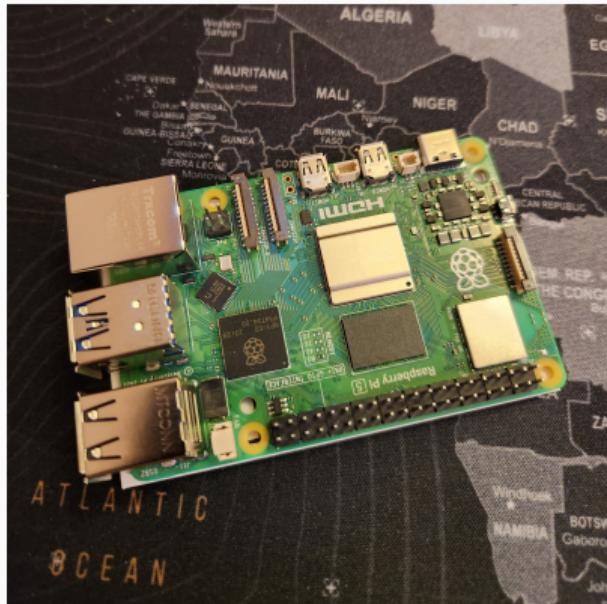
➤ Utilizamos Perspective (BERT family) para clasificar toxicidad, insultos y amenazas:

Datos de protesta

Datos de la Convención

Algoritmo adicional ToxicBERT para datos de protesta

Datos y clasificación de toxicidad



Los datos de las protestas se clasificaron completamente de forma local en una **Raspberry Pi 5**. El 41% de los lotes de datos de la Convención se clasificaron allí y el resto en la nube.

Esto nos permitió reducir la huella de carbono de nuestras tareas computacionales en $\sim 39\%$.

* En kg de CO₂ es casi 54 cortes de vacuno de 300g.

** Fotografía de la Raspberry Pi 5 usada para la clasificación.

Etiquetado humano

Etiquetado humano

El etiquetado será en  Labelbox a través de sus cuentas personales. Demostración después de esta presentación.

- Pool de 1.000 mensajes (balanceados por quintiles de toxicidad) por cada conjunto
- Umbral de dos mensajes por codificador (pool total 4.000 observaciones)
- Máximo posible por codificador de 2.000

Clasifique la categoría del comentario como TÓXICO o NO TÓXICO

TÓXICO: Comentarios groseros, irrespetuosos o poco razonables que probablemente hagan que alguien abandone la conversación o deje de compartir su punto de vista.

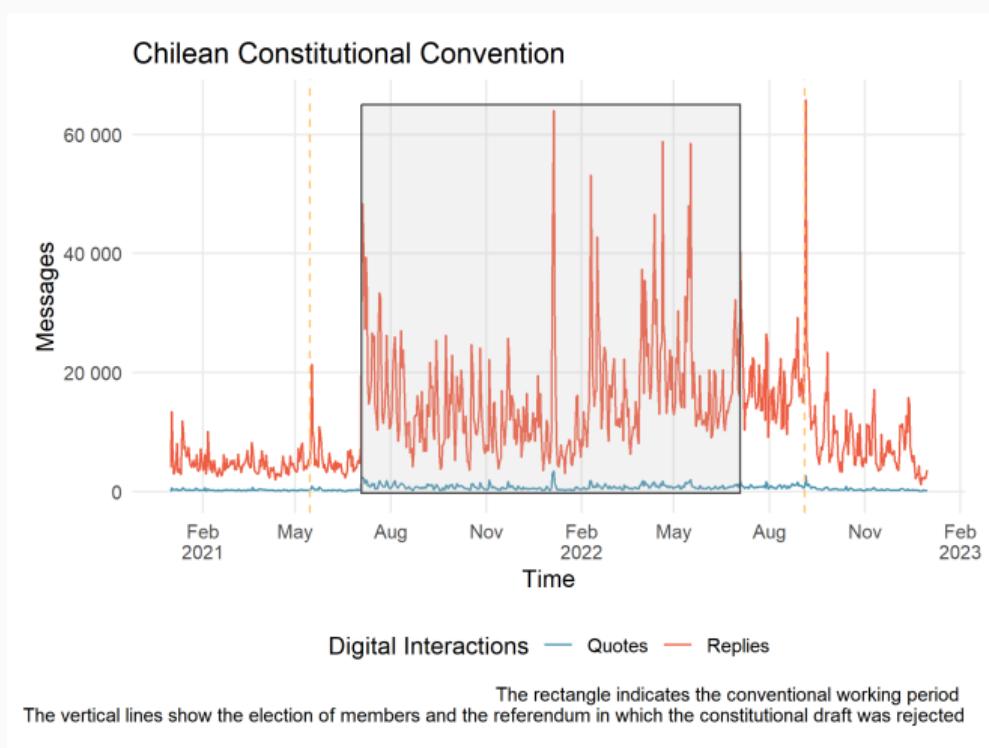
NO TÓXICO: Comentarios civiles o amables que probablemente no desalienten la conversación.

Trabajo futuro

Paper: Digital Incivility during the Constitutional Convention

 B. González-
Bustamante & S.
Rivera

Dinámicas de toxicidad e incivilidad durante la Convención y el efecto de *shocks* informativos.



OpenAI Project: LLMs, Toxicity and Mass Protests



👤 B. González-Bustamante

Benchmarking piloto con diferentes LLMs, el listado (aún por acotar):

llama2, llama3, gemma, aya, dolphin-llama3, orca2, qwen, qwen2, mistral-openorca, nous-hermes2, nous-hermes2-mixtra, wizard-vicuna

Comentarios finales

Comentarios finales

- Los datos estarán disponibles para otros usos.
-  El [repositorio de GitHub](#) ya está habilitado.
-  Pueden revisar el [estatus de Labelbox](#) si tienen problemas con la plataforma.
-  Tenemos esta [Shiny App prototipo](#) con los datos de la Convención (por favor no difundir de momento).

Muchas gracias

Cualquier pregunta, me pueden escribir

b.a.gonzalez.bustamante@fgga.leidenuniv.nl

bgonzalezbustamante.com